



# Time-Course Gene Set Analysis for Longitudinal Gene Expression Data

Boris P. Hejblum, Jason Skinner, Rodolphe Thiébaut

## ► To cite this version:

Boris P. Hejblum, Jason Skinner, Rodolphe Thiébaut. Time-Course Gene Set Analysis for Longitudinal Gene Expression Data. PLoS Computational Biology, 2015, 11 (6), pp.e1004310. 10.1371/journal.pcbi.1004310 . hal-01203446

**HAL Id: hal-01203446**

**<https://inria.hal.science/hal-01203446>**

Submitted on 13 Nov 2016

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

RESEARCH ARTICLE

# Time-Course Gene Set Analysis for Longitudinal Gene Expression Data

Boris P. Hejblum<sup>1,2,3,4,5</sup>, Jason Skinner<sup>4,5</sup>, Rodolphe Thiébaut<sup>1,2,3,4,5\*</sup>

**1** Univ. Bordeaux, ISPED, Centre INSERM U897-Epidemiologie-Biostatistique, F-33000 Bordeaux, France, **2** INSERM, ISPED, Centre INSERM U897-Epidemiologie-Biostatistique, F-33000 Bordeaux, France, **3** INRIA, Team SISTM, F-33000 Bordeaux, France, **4** Vaccine Research Institute-VRI, Hôpital Henri Mondor, Créteil, France, **5** Baylor Institute for Immunology Research, Dallas, Texas, United States of America

\* [rodolphe.thiebaut@u-bordeaux.fr](mailto:rodolphe.thiebaut@u-bordeaux.fr)



## OPEN ACCESS

**Citation:** Hejblum BP, Skinner J, Thiébaut R (2015) Time-Course Gene Set Analysis for Longitudinal Gene Expression Data. PLoS Comput Biol 11(6): e1004310. doi:10.1371/journal.pcbi.1004310

**Editor:** Quaid Morris, University of Toronto, CANADA

**Received:** November 18, 2014

**Accepted:** April 30, 2015

**Published:** June 25, 2015

**Copyright:** © 2015 Hejblum et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** The data sets supporting the results of this article are available in the GEO (Gene Expression Omnibus) repository. The gene expression data of the DALIA-1 trial are available under the following identifier: GSE46734 (<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE46734>). The data from the vaccine study of Obermoser et al. are available under the following identifier: GSE30101 (<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE30101>).

**Funding:** This work was supported by the Investissements d'Avenir program managed by the ANR under reference ANR-10-LABX-77 and by the Vaccine Research Institute (VRI), F-94010 Creteil, France. BPH is recipient of a Ph.D. fellowship from

## Abstract

Gene set analysis methods, which consider predefined groups of genes in the analysis of genomic data, have been successfully applied for analyzing gene expression data in cross-sectional studies. The time-course gene set analysis (TcGSA) introduced here is an extension of gene set analysis to longitudinal data. The proposed method relies on random effects modeling with maximum likelihood estimates. It allows to use all available repeated measurements while dealing with unbalanced data due to missing at random (MAR) measurements. TcGSA is a hypothesis driven method that identifies a priori defined gene sets with significant expression variations over time, taking into account the potential heterogeneity of expression within gene sets. When biological conditions are compared, the method indicates if the time patterns of gene sets significantly differ according to these conditions. The interest of the method is illustrated by its application to two real life datasets: an HIV therapeutic vaccine trial (DALIA-1 trial), and data from a recent study on influenza and pneumococcal vaccines. In the DALIA-1 trial TcGSA revealed a significant change in gene expression over time within 69 gene sets during vaccination, while a standard univariate individual gene analysis corrected for multiple testing as well as a standard a Gene Set Enrichment Analysis (GSEA) for time series both failed to detect any significant pattern change over time. When applied to the second illustrative data set, TcGSA allowed the identification of 4 gene sets finally found to be linked with the influenza vaccine too although they were found to be associated to the pneumococcal vaccine only in previous analyses. In our simulation study TcGSA exhibits good statistical properties, and an increased power compared to other approaches for analyzing time-course expression patterns of gene sets. The method is made available for the community through an R package.

## Author Summary

Gene set analysis methods use prior biological knowledge to analyze gene expression data. This prior knowledge takes the form of predefined groups of genes, linked through their biological function. Gene set analysis methods have been successfully applied in

the Ecole des Hautes Etudes en Santé Publique (EHESP) Doctoral Network. The DALIA-1 study was supported by a grant from the French National Agency for Research on AIDS and Viral Hepatitis (ANRS: L'Agence Nationale pour la Recherche contre le SIDA et les hépatites virales). The VRI participated in study design, data collection and analysis for the DALIA trial. The funders had no role in decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

transversal studies, their results being more sensitive and interpretable than those of methods investigating genomic data one gene at a time. The time-course gene set analysis (TcGSA) introduced here is an extension of such gene set analysis to longitudinal data. This method identifies *a priori* defined groups of genes whose expression is not stable over time, taking into account the potential heterogeneity between patients and between genes. When biological conditions are compared, it identifies the gene sets that have different expression dynamics according to these conditions. Data from 2 studies are analyzed: data from an HIV therapeutic vaccine trial, and data from a recent study on influenza and pneumococcal vaccines. In both cases, TcGSA provided new insights compared to standard approaches thanks to an increased sensitivity compared to other approaches. Those results highlight the benefits of the TcGSA method for analyzing gene expression dynamics.

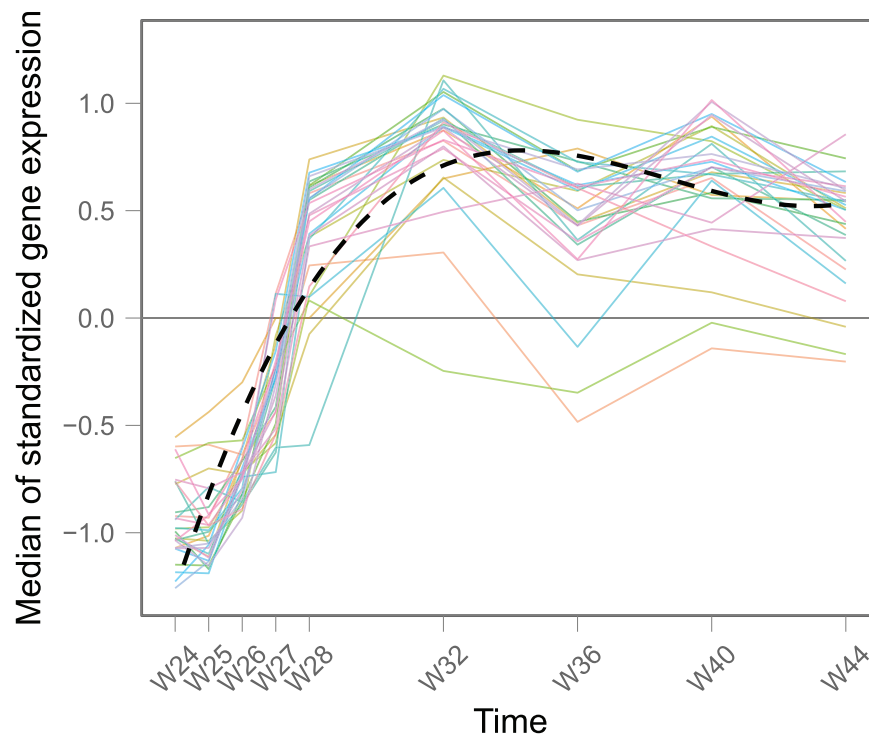
This is a *PLOS Computational Biology* Methods paper.

## Introduction

Microarray experiments are increasingly used for evaluating changes in gene expression over time. The analysis of the temporal change of gene expression should help in understanding the complex mechanisms of gene regulation. For instance, transcriptional profiles have been repeatedly measured to study the change in gene expression during the natural history of SIV/HIV infection [1, 2] or to evaluate the effect of vaccines [3, 4]. In the applications considered in this paper [5, 6], the investigators wanted to detect the genes for which the abundance changed over time after a vaccination (against HIV, influenza or pneumococcus).

In order to analyze such longitudinal high-dimensional data, several approaches have been suggested including a gene-by-gene statistical analysis [7, 8], dimension reduction methods [9] or gene set analysis [10]. A gene set is a group of genes that are *a priori* co-regulated or functionally linked. Examples of such gene set relating to biological processes or pathways are those defined by KEGG [11], Gene Ontology [12] or Chaussabel's functional modules [13]. The gene set analysis [14–16] is supposed to be more powerful than a gene-by-gene analysis because it can detect a change of expression of a group of genes although none of them show a very high absolute fold change. Furthermore, a change of all genes in a given pathway may be biologically more meaningful than a large increase of a single gene. Also, provided that the gene sets are well defined, the result should be more sound and comparable across studies than a gene-by-gene analysis [14]. Finally, gene set analysis avoids a second step for a global interpretation as described in the “bottom up” approach [10, 17].

The analysis of longitudinal microarray experiments through a gene set approach is not trivial because the dynamics of gene expressions inside a gene set can be complex and heterogeneous. This has already been underlined in some of the approaches developed to analyze gene sets [15, 18–20]. Fig 1 shows an example of a homogeneous gene set, whereas Fig 2 shows an example of a heterogeneous one. Actually, such a heterogeneity is frequently observed [20], and cannot be ignored, as genes inside a functional gene set are not expected to change their expression synchronously. Moreover this heterogeneity can be biologically meaningful by itself. Prieto et al. [21] provide an example from a cancer application, where deregulated pathways

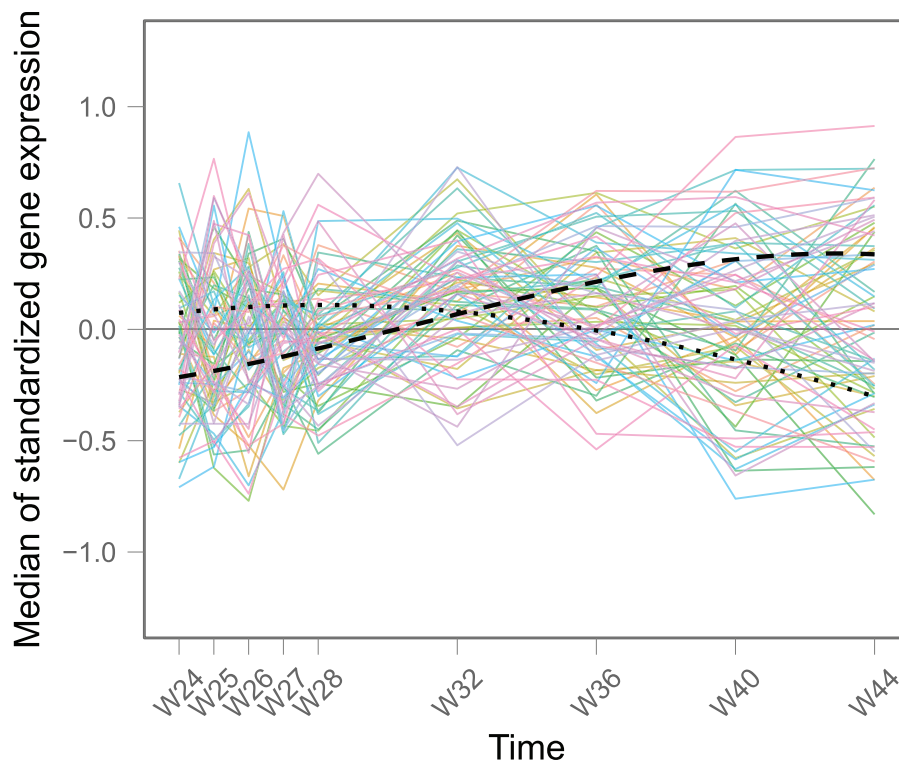


**Fig 1. Example of a homogeneous gene set (M1.2: interferon—from the DALIA trial, after treatment interruption).** Each line is the median expression of a gene inside this particular gene set across all the patients. The expression of the genes inside this gene set is quite homogeneous and it is easy to identify a global time trend, displayed by the dashed black line (smoothed median). For more information see the presentation of the DALIA-trial in the “Results” section.

doi:10.1371/journal.pcbi.1004310.g001

are of primary biological interest. They identified heterogeneous gene sets linked to acute promyelocytic leukemia. Another example is given by Hu et al.: pathways affected by the HER2, such as the KEGG pathways of ‘Ubiquitin mediated proteolysis’, ‘Glioma’, and ‘Prostate cancer’ were identified by studying heterogeneity [22]. The main advantage of detecting the heterogeneity inside a gene set is to detect any change over time whatever the specification of the model for the trends. In other words, the dynamics of gene expression inside a stable gene set will be summarized by a flat slope and no heterogeneity. Hence, in the spirit of [19], to find any significant change of the overall expression of genes inside a gene set over time, we suggest to look for any significant trend over time or any heterogeneity between gene trends inside the gene set.

Several approaches have already been proposed to analyze longitudinal measurements of gene expression [7, 23–28], but only a few include gene set analysis [10, 18, 29–31]. Among the latter, all but Nueda et al. [18] fail to account for possible heterogeneity inside a gene set. An extension of the popular Gene Set Enrichment Analysis (GSEA) method [14] is available for the analysis of time series data: GSEA for time series. Unfortunately, it does not account for the structure of longitudinal data, simply treating all observations as independent and calculating Pearson correlation of the genes with the time variable. Therefore this is quite a limited modeling strategy for longitudinal data. The globalANCOVA procedure developed by Hummel et al. [29] focuses on the comparison of groups, testing whether there is a group influence on change



**Fig 2. Example of a heterogeneous gene set (M4.16: cell cycle—from the DALIA trial, after treatment interruption).** Each line is the median expression of a gene inside this particular gene set across all the patients. The expression of the genes inside this gene set is rather heterogeneous. This makes it difficult to identify any time trend, as the mean expression inside this gene set stays close to zero. However a closer look reveals two distinct time trends, displayed by the two respectively dashed and dotted black lines (smoothed medians). For more information see the presentation of the DALIA-trial in the “Results” section.

doi:10.1371/journal.pcbi.1004310.g002

over time of any gene expression inside a gene set. In practice, the global null hypothesis tested is quite flexible relying on the ANOVA framework, but cannot accommodate missing values. Wang et al. [10] proposed to use a linear mixed effects model to explain gene expression inside a gene set. They considered a random effect for the array level rather than for the patient or the gene level. Zhang et al. [30] proposed a robust non-parametric approach to compare gene expression dynamics between different treatment-groups. Wu et al. [31] proposed the CAMERA procedure, a gene set test based on linear modeling that takes into account inter-gene correlations. However, CAMERA accounts for neither potential heterogeneity inside a gene set nor for repeated measurement correlation. Of note, it is not possible to look at the change of gene expression in only one group using Zhang et al. [30] approach. In contrast, the PCA-maSigFun procedure developed by Nueda et al. [18, 32] can account for possible heterogeneity inside a gene set. It is based on a Principal Component Analysis (PCA) of each gene set followed by a linear regression of the significant principal components (i.e. components that have a variation above the mean gene variance) over time. However, they did not consider time-course experiments where repeated measures are available for multiple patient.

Gene set analysis methods can also be distinguished by their choice of the null hypothesis. Those can be classified into two main types of hypothesis: i) the competitive null hypothesis, that tests the genes inside a given gene set against all the other genes outside the gene set; ii) the

self-contained null hypothesis, that only uses the genes inside the gene set of interest [20, 33, 34]. In the present paper, interest is focused on self-contained null hypotheses because the question was “Which gene sets have a change of gene abundance over time?”. According to Emmert-Streib et al. [35], self-contained gene set tests are biologically easier to interpret and can be more powerful compared to competitive tests. Self-contained gene set tests are especially appropriate in a hypothesis driven context where *a priori* defined gene sets are validated and relevant regarding the biological question.

We propose the implementation of a hypothesis driven method that directly tests the time-course significance of predefined gene sets: the *Time-course Gene Set Analysis* (TcGSA). It relies on the use of linear mixed effect models, a very useful and well established statistical tool [36, 37] especially suited for longitudinal settings. By using all available repeated measures, it provides increased statistical power. TcGSA can accommodate for heterogeneity of gene expression within the gene sets through random effects, and is robust to unbalanced designed due to missing (at random) values thanks to the maximum likelihood estimates. No previously proposed approach combines all of TcGSA features. A simulation study demonstrated the good statistical performance of the proposed method. It has been applied to two studies: one HIV vaccine trial, and one influenza and pneumococcal vaccine study [6], using the same definition of gene sets [13] that is increasingly used in systems immunology research [38–42]. Compared to gene-by-gene analyses, TcGSA disclosed changes of additional gene sets that endorse previous conclusions [6], but also revealed common pathways across the three vaccines.

## Methods

### Time-course gene set analysis

The TcGSA method includes three steps: 1) modeling gene expression in a gene set with mixed models, 2) testing the significance of a gene set, and 3) estimating individual gene profiles.

**1. Modeling gene expression in a gene set with mixed models.** Let  $S$  be a gene set of interest. We start by the case of a one group experiment, where each patient act as her/his own respective control, her/his condition changing over time. The expression of genes inside  $S$  is modeled over time according to a function  $f$  as:

for all the genes  $g \in S$ ,

$$y_{gpi} = \mu + \beta_g + c_{gp} + f_g(t_i) + \varepsilon_{gpi} \quad (1)$$

where  $y_{gpi}$  is expression of the  $g^{th}$  gene for the  $p^{th}$  patient at the  $i^{th}$  time,  $\mu$  is the intercept in the gene set  $S$ ,  $\beta_g$  is the fixed effect of the  $g^{th}$  gene,  $c_{gp} \sim \mathcal{N}(0, \sigma_c)$  is a random effect grouped by the  $g^{th}$  gene of the  $p^{th}$  patient,  $t_i$  is the  $i^{th}$  measurement time,  $\varepsilon_{gpi} \sim \mathcal{N}(0, \sigma)$  is an error term. Finally  $f_g(t_i)$  is a function of time, that can be linear, polynomials, etc. Every time coefficient of the trend  $f_g(t_i)$  is actually divided into a fixed effect  $\eta$ . (representing the average trend in the gene set  $S$ ) and a random effect  $h_{g,\cdot} \sim \mathcal{N}(0, \sigma_h)$  grouped on the gene  $g$ , accounting for the possible heterogeneity between the genes in the gene set  $S$ . In this paper we focus on three forms for  $f_g$  (but other forms, such as exponential, etc. could easily be envisaged):

- linear polynomials:

$$f_g(t) = (\eta_1 + h_{g,1})t$$

- cubic polynomials:

$$f_g(t) = (\eta_1 + h_{g,1})t + (\eta_2 + h_{g,2})t^2 + (\eta_3 + h_{g,3})t^3$$



- natural cubic splines:

$$f_g(t) = \sum_{k=1}^{K+1} (\eta_k + h_{g,k}) N_k(t)$$

where the  $N_k(t)$  form the natural cubic splines basis [43] for the time variable  $t$  (with  $K$  internal knots),  $\eta_k$  are the fixed effects of time shared across the gene set  $S$ , and  $h_{g,k}$  are the random effects of time accounting for possible heterogeneity between genes.  $(h_{g,1}, \dots, h_{g,d}) \sim \mathcal{N}(0, \Sigma_h)$  with  $d$  being the degree of the time function, and for  $k = 1, \dots, d$   $h_{g,k} \sim \mathcal{N}(0, \sigma_{h_k})$ . Alternatively, one can make the assumption that the patient effect is the same for all the genes. In that case, the random effect  $c$  is no longer grouped on the gene level, and the model can be written as:

for all the genes  $g \in S$ ,

$$y_{gpi} = \mu + \beta'_g + c'_p + f_g(t_i) + \varepsilon_{gpi} \quad (1bis)$$

with  $c'_p \sim \mathcal{N}(0, \sigma_{c'})$  the random effect of the patient  $p$ , and  $\beta'_g \sim \mathcal{N}(0, \sigma_{\beta'})$  the random effect of the gene  $g$ . This alternative modeling has the advantage to be more parsimonious than the Eq (1), with less parameters to be estimated.

Let's now consider the case of a multiple group experiment (such as treatment/vaccine groups for instance). The expression of genes inside  $S$  is modeled over time according to a function  $f_{g,m}$  that is now stratified on the groups:

for all the genes  $g \in S$ ,

$$y_{mgpi} = \mu + \beta_g + \delta_m + c_{gp} + f_{g,m}(t_i) + \varepsilon_{gpi} \quad (2)$$

where  $m$  indicates which group is concerned and  $\delta_m$  is the fixed intercept of the  $m^{th}$  group, everything else being the same as in the Eq (1).

**2. Testing the significance of a gene set.** In TcGSA, a “significant” gene set is a gene set whose expression is not stable either over time (in one group experiments) or over groups (in several groups experiments), once between genes and patients variability is taken into account. In other words, we want to test the significance of the time trend while being sensitive to both homogeneous and heterogeneous changes of gene expression over time inside a gene set. Testing the significance of a given gene set  $S$  therefore means testing both fixed and random effects at once, in a single test. A likelihood ratio test is the natural way to do so, fitting models under both the null hypothesis and the alternative.

In the case of one group experiment (Eqs (1) and (1bis)) the null hypothesis ( $H_0$ ) is that the genes inside  $S$  are stable over time, *i.e.* that their expressions are constant and homogeneous over time (all coefficients of the function of time  $f$  are not significantly different from zero). The alternative hypothesis ( $H_1$ ) is that the genes inside  $S$  vary significantly over time:

$$(H_0) : \quad \forall k, \eta_k = 0 \quad \text{and} \quad \sigma_{h_k} = 0 \quad (1.0)$$

$$(H_1) : \quad \exists k, \eta_k \neq 0 \quad \text{or} \quad \sigma_{h_k} \neq 0 \quad (1.1)$$

In the case of a multiple group experiment (Eq (2)), the null hypothesis is that inside the gene set  $S$ , the evolution of gene expressions over time is the same regardless of the group. The alternative hypothesis is that time trends  $f$  are different depending on the group  $m$ :

$$(H_0) : \quad \forall m, f_{g,m}(\cdot) = f_g(\cdot) \quad (2.0)$$

$$(H_1) : \quad \exists m, m' \text{ such that } f_{g,m}(\cdot) \neq f_{g,m'}(\cdot) \quad (2.1)$$

In both case, one model is fitted under the null hypothesis, and one is fitted under the alternative. The likelihood ratio is then computed.

However, since both fixed and random effects are tested simultaneously in this likelihood ratio, its null distribution is not the standard chi-square distribution (because of boundary constraints due to the variance of random effects). According to Self et al. [44], it can be approximated by a mixture of chi-square distributions with the following formula:

$$LR_{H_0} \sim \sum_{k=q}^{q+r} \binom{r}{k-q} 2^{-r} \chi_{(k)}^2$$

where  $q$  is the number of fixed effects and  $r$  the number of random effects to be tested simultaneously. This approximation implies that the tested random effects are independent of one another [45–47]. This seems an acceptable assumption according to our simulations under the null (see S1 and S2 Figs). This allows to compute a p-value for the significance of the variation of a given gene set over time.

When several gene sets are investigated at a time, it is necessary to take into account multiple testing. A number of procedures are available to do so [48]. As the TcGSA is mostly an exploratory analysis (even though hypothesis driven in the sense that gene set are defined *a priori*), we recommend using the Benjamini-Yekutieli procedure for controlling the False Discovery Rate [49], as gene sets are necessarily correlated between each others and this procedure is robust to some of these dependances. Other multiple testing correction procedures are available in the TcGSA R package.

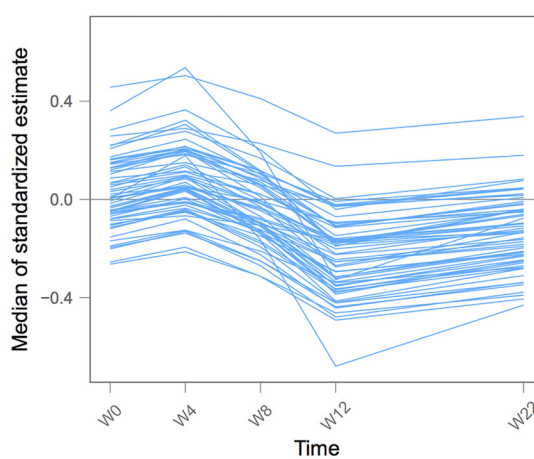
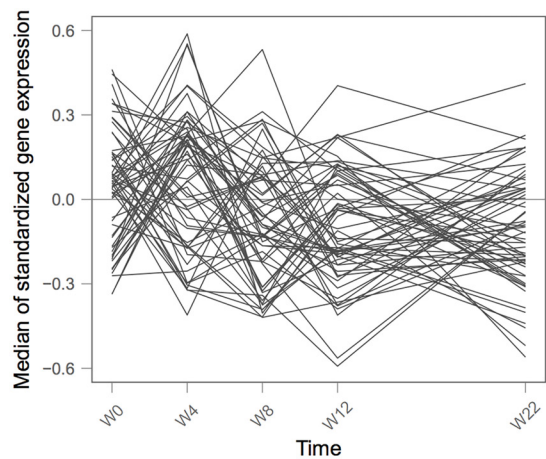
**3. Estimation of individual gene profiles.** In the estimation of linear mixed model, it is common to use the Restricted Maximum Likelihood (REML) instead of the classic Maximum Likelihood (ML) in order to avoid biased estimates of the variance components [50]. But note that REML cannot be used to estimate the likelihood ratios presented here. Indeed, REML estimation of the likelihood ratio between two models can only be used when both models have the same fixed part [51]. Since here the compared models (under  $H_0$  and under  $H_1$ ) have different fixed components (due to the  $\eta$  coefficients under  $H_1$ ), the use of ML estimation is needed.

For the inference of the random effects, Best Linear Unbiased Predictor (BLUP) are used [52], giving access to estimations of a single profile for each gene among a gene set, in each patient. As a result, the estimations from the mixed model are shrunk towards the average expression inside the gene set. This shrinkage occurs when the residuals variability is relatively large compared to the the random effects estimated variances [52]. The mixed model uses the repeated pattern of the longitudinal measurements to structure the variation. Its estimations give smoother trajectories for the genes than the raw data, which makes the general evolution of the set clearer [53], as it can be seen in Figs 3 and 4.

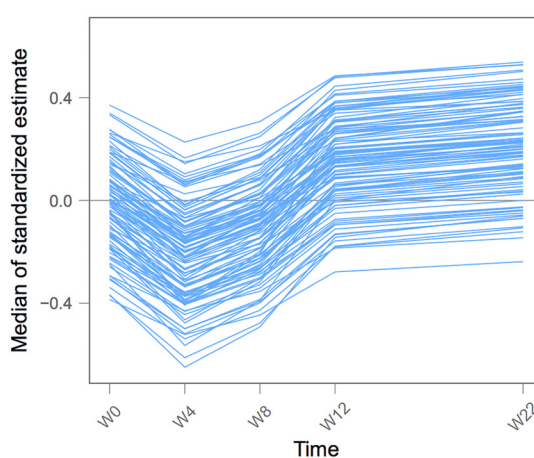
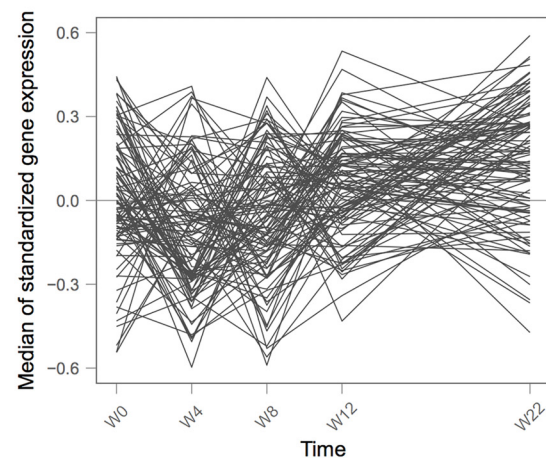
## Characterization and visualization of dynamics

**Dynamic of a significant gene set.** Once a gene set  $S$  has been identified as significant through the previous mixed likelihood ratio statistics, a summary of its dynamic over time is needed. However, due to the possible heterogeneity of  $S$ , giving a summary representation of  $S$  dynamic is not obvious. We propose to automatically identify the number of trends in a significant gene set from the fit of the model. Predicted gene expressions from the linear mixed model are clustered, and the optimal number of trends is selected with the gap statistic [54]. It is a formalization of the elbow criterion for the within-cluster variance. In order to determine the optimal partition of each gene set here, the gap statistics is applied onto a hierarchical clustering of gene expressions inside each gene set. Then the median within each of the identified clusters can summarize each trend. Therefore, gene sets are actually split when heterogeneous,

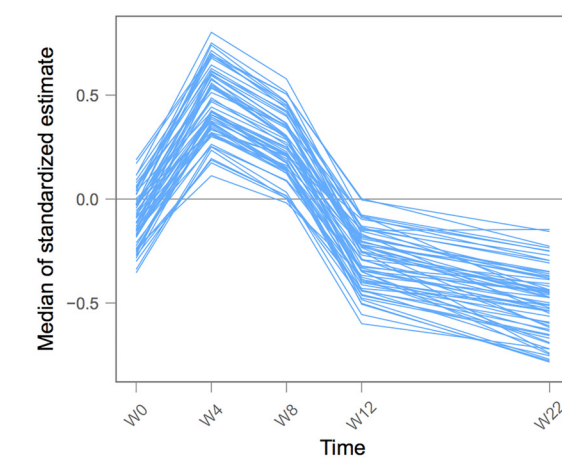
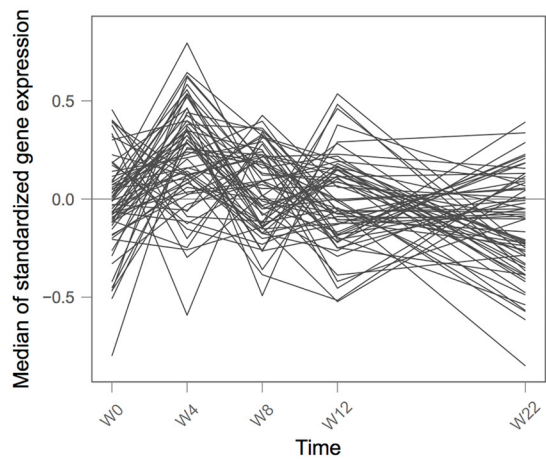




**M4.1: T Cell**



**M4.6: Inflammation**



**M6.7: B Cell**

**Fig 3. Three significant gene sets during pre-ATI in DALIA-1.** Each line is the median over the patients of the expression of one gene. Each graph shows all the genes in one particular gene set. The left graph displays the raw gene expression, the right one displays the estimations from the mixed model for the same gene set. The expressions have been centered and reduced for this representation.

doi:10.1371/journal.pcbi.1004310.g003

before being summarized. The predicted gene expression from the linear mixed model is used for this (and not the observed expression) because smoothness of trajectories facilitates classification [53]. Examples of such representations are given in Figs 3 and 4.

**Global dynamics.** Most often, TcGSA will be used to investigate a large number of gene sets (from a few hundreds to a few thousands). This multiplicity can make visualization of the results more challenging, in addition of requiring a multiple testing correction. TcGSA is designed to identify gene sets that shows a simultaneous evolution of gene expression, but possibly of a small intensity. The method can therefore be quite sensitive, and it can be of interest to rank the significant gene sets to identify the most acute signals. The likelihood ratio provides insight on the magnitude of the variation of each gene set. The percentile of their corresponding likelihood ratio gives an idea of the importance of the variation for a significant gene set. Examples of such representations are given in Fig 5.

## Implementation

The TcGSA method has been implemented in R as a package *TcGSA*, whose latest release is available from the CRAN repository (<http://cran.r-project.org/web/packages/TcGSA/index.html>).

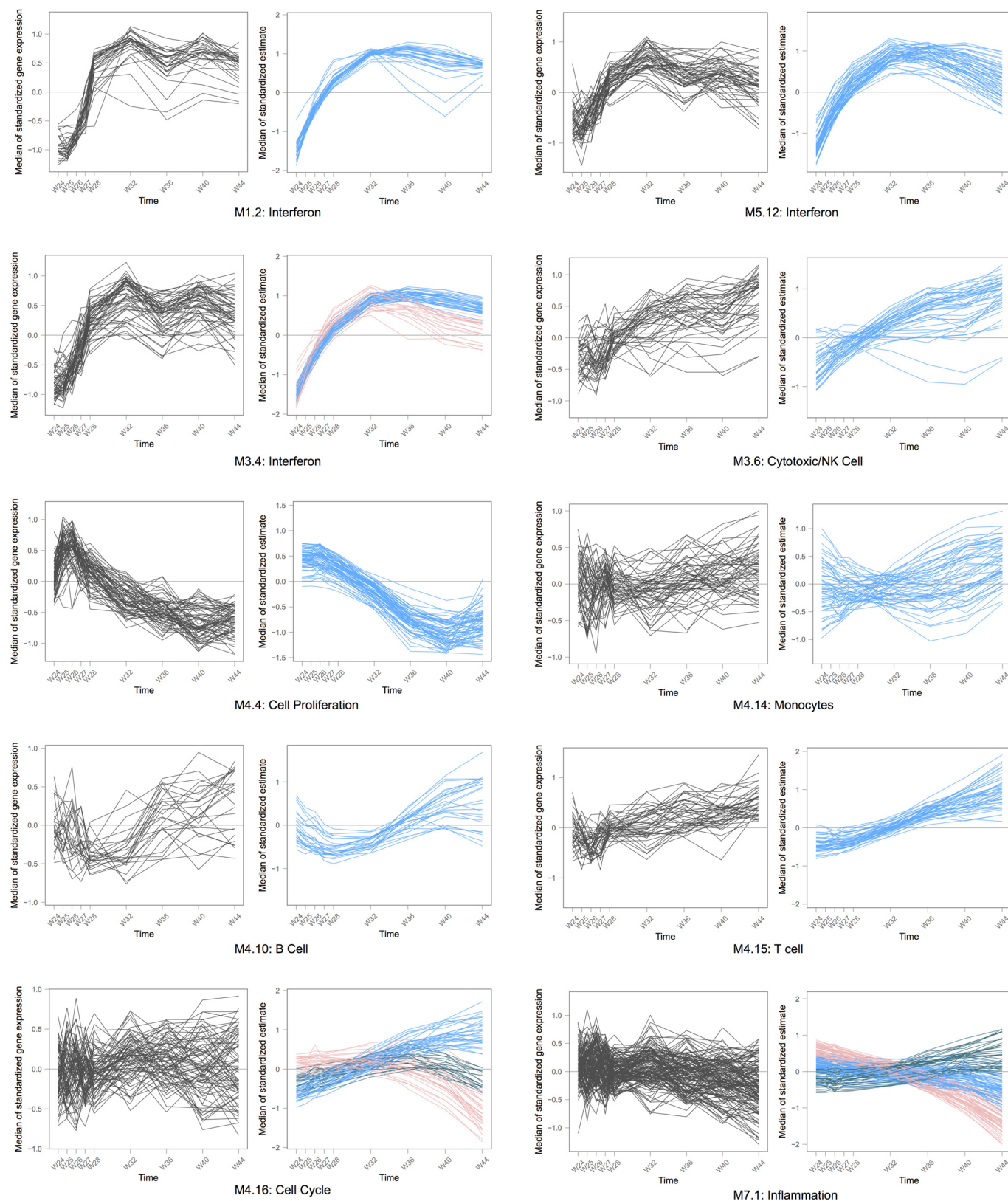
## Results

### Motivating example: The DALIA-1 trial

The DALIA-1 trial is a phase 1 therapeutic HIV vaccine trial whose details are described on <http://clinicaltrials.gov> (ClinicalTrial.gov identifier: NCT00796770) and in [5]. The vaccine candidate was based on ex-vivo generated interferon- $\alpha$  dendritic cells loaded with HIV-1 lipopeptides and activated with lipopolysaccharide. The objectives of the trial were to evaluate the safety of the strategy and to evaluate the immune response to the vaccine. For the purpose of the present paper, we focus on the gene expression component of this study. Gene abundance in whole blood was measured through Illumina HumanHT-12 v4 Expression BeadChips.

**The DALIA-1 trial design.** All of the nineteen HIV infected patients received the therapeutic vaccine while under antiretroviral treatment. The patients received four injections at week 0, 4, 8 and 12. This vaccination period was followed by an antiretroviral treatment interruption (ATI) at week 24. The patients were followed up to week 48. Antiretroviral treatment was resumed from week 24 to week 48 at any time under the following criteria: i) if the patients or their doctors wished so; ii) if CD4+ T-cell count was  $< 350$  cells/ $\mu$ L and  $< 25\%$  of total lymphocytes. Fourteen time points (five in pre-ATI from week 0 to week 22, and nine in post-ATI from week 24 to week 44) were used for this analysis (see Fig 6). One patient was removed from the analysis as his/her antiretroviral treatment compliance was irregular during the vaccination phase.

In the following analysis, two distinct datasets were considered: pre-ATI and post-ATI. The two datasets were normalized separately—via a normal-exponential convolution model [55, 56], followed by the application of the *ComBat* method [57] to correct for batch effects.



**Fig 4. Ten significant gene sets during post-ATI in DALIA-1.** Each line is the median over the patients of the expression of one gene. Each graph shows all the genes in one particular gene set. The left graph displays the raw gene expression, the right one displays the estimations from the mixed model for the



same gene set. If several dynamics are identified by the gap statistics among the estimated expressions inside one gene set, they are displayed in different colors—such as for the gene sets M 4.16 and M 7.1 that each features three different dynamics. The expressions have been centered and reduced for this representation.

doi:10.1371/journal.pcbi.1004310.g004

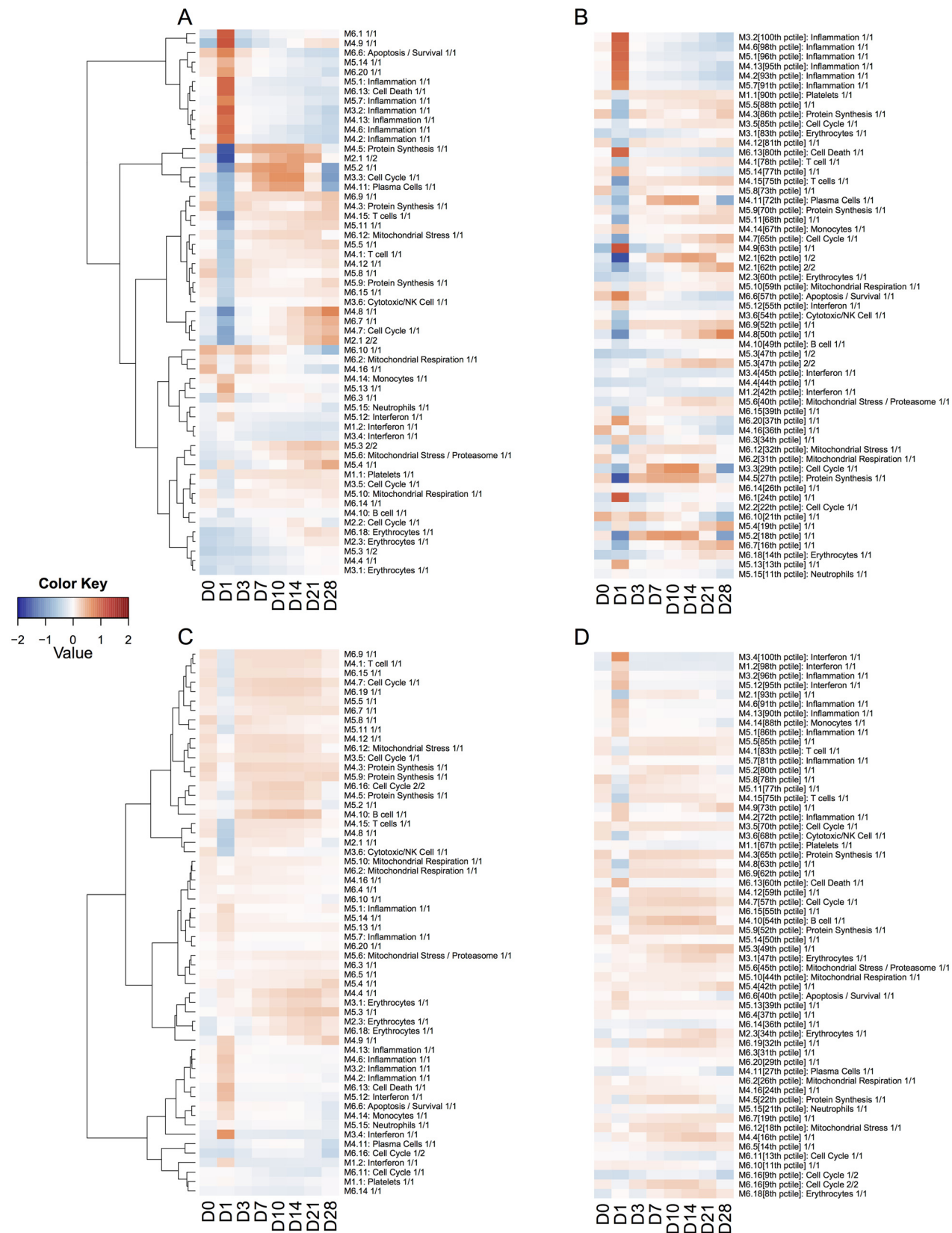
Splitting the data allows us to study separately the vaccine effect and the treatment interruption, otherwise the ATI effect would mask any noticeable vaccine effect, because of the huge modification of gene expression related to viral replication [1, 2]. We investigated the gene sets defined by Chaussabel et al. [13], which are oriented towards the immune system. The definition and annotations of those 260 gene sets (called ‘Modules’) are available online ([http://www.biir.net/public\\_wikis/module\\_annotation/V2\\_Trial\\_8\\_Modules](http://www.biir.net/public_wikis/module_annotation/V2_Trial_8_Modules)).

**Pre-ATI: The vaccination phase.** During the vaccination period, a standard gene-by-gene mixed model analysis, with a cubic polynomial function of time, did not found any significant change of gene abundance at a 5% False Discovery Rate (see Table 1). However, during this vaccination phase, cytokines production analysis of the same blood samples (as measured by Luminex or intracellular staining) have showed that a response was induced by the vaccine at week 16 [5]. Therefore, one expected to observe a signal at the gene expression level between week 0 and week 16, the gene expression preceding molecular activation. Although the measurements were not performed in the hours or days following vaccination, the changes of gene abundance may reflect a change of the equilibrium of the overall expression in some gene sets. This kind of results has already been reported in cross-sectional studies [58]. Likewise, GSEA for time-series did not identify any significant gene set during vaccination. This can be explained by the lack of power of GSEA for time-series, as this method does not take into account the repeated structure of the data and is not suitable for longitudinal measurements. Finally, CAMERA did not identified any significant gene set either, in spite of testing a competitive null hypothesis.

We applied the Time Course Gene set Analysis (see Methods) that allows to detect any change over time of gene abundance inside a gene set by detecting either trends over time or heterogeneity between gene dynamics. Fitting the Eq (1) with a cubic polynomial function of time, 69 gene sets out of 260 turned out to vary significantly. Fig 3 displays the raw sample-normalized and batch-corrected) and estimated gene expressions of 3 of the significant gene sets identified by TcGSA: T-cell, inflammation and B-cell gene sets. The identification of gene sets such as M4.1: T-cell (that includes CD402, CCR7, BCL2) was expected with regards to the CD4 T-cell response observed at Week 16 [5]. Also, the gene sets M4.6: inflammation and M6.7: B-cell are good examples of how smoothing from the estimations can give a much clearer dynamic pattern compared to the raw expression (see Methods).

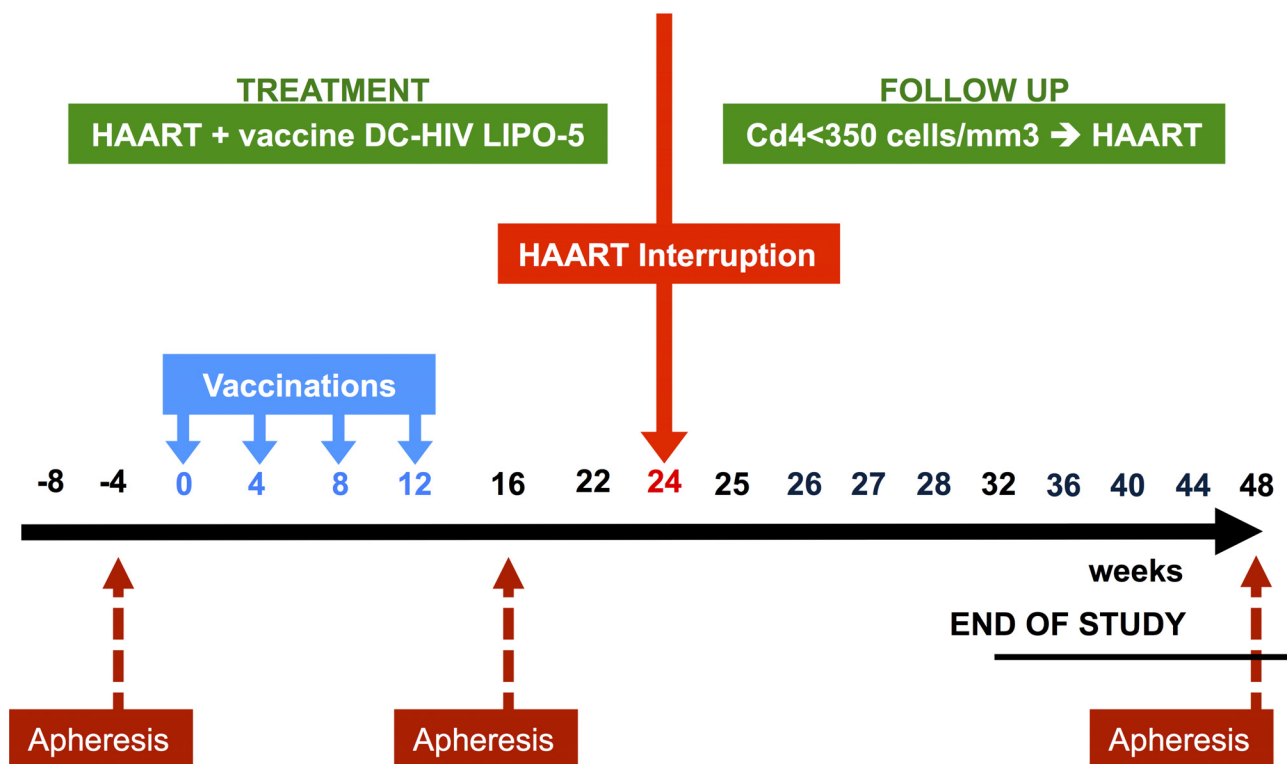
**Post-ATI: After antiretroviral treatment interruption.** Eq (1) was then fitted to the data after antiretroviral treatment interruption that occurred at week 24: 216 gene sets out of 260 were found to be significant with a cubic polynomial function of time. Fig 4 displays the raw and estimated expressions of nine of those significant gene sets. It features heterogeneous gene sets, such as M4.16 and M7.1, which are both also good examples of the shrinkage that occurs with the estimations (see Methods). Meanwhile, GSEA for time series identifies 67 significant, whereas CAMERA identified only 2. These results are consistent with the lower statistical power found in our simulation study.

The large number of significant gene sets post-ATI illustrates the tremendous impact of the treatment interruption on the organism. Followed by a viral rebound, the treatment interruption is indeed a major event that triggers the expression of thousand of genes. Indeed, a gene-by-gene analysis revealed 3,389 significant probes (more than 10% of all the investigated



**Fig 5. Heatmap of estimated dynamics from the significant gene sets among the 62 investigated gene sets when comparing vaccines arms to placebo arm.** The median estimated gene expression over the patients is used for each trend. Each trend has seen its values reduced (so that its variance is 1) in order to make the dynamics more comparable. Each row is a group of gene having the same trend inside a gene set, and each column is a time point. The color key represents the median of the standardized estimation of gene expression over the patients for a given trend in a significant gene set. It becomes red as median expression is up-regulated or blue as it is down-regulated compared to the value in the placebo (saline) at the same time. A and C show the hierarchically clustered trends for pneumococcal and flu respectively. B and D show the same trends but instead ranked by decreasing likelihood ratio percentiles of the associated gene set, for pneumococcal and flu respectively.

doi:10.1371/journal.pcbi.1004310.g005



**Fig 6. DALIA-1 trial design.** Gene expression was measured at each time point, represented by a week number above the time axis. The trial was composed of two separated stages: (1) the treatment phase, during which the patients were vaccinated but remained under antiretroviral treatment; and (2) the follow-up phase commencing after the week 24 antiretroviral treatment interruption. Those two phases will be referred as pre-ATI and post-ATI respectively. The three apheresis time points were removed from the analysis due to a possible effect of the apheresis on the gene expression samples, and so was the first measurement (week -8) occurring right at the inclusion in the study.

doi:10.1371/journal.pcbi.1004310.g006

probes—an unusually high number of differentially expressed genes). The immune system is very much in demand during the viral rebound. Therefore most of the gene sets from the Modules defined by Chaussabel et al. [13] are activated, as they are tightly linked with the immune system activity. Of particular interest are the three gene sets M1.2, M3.4 and M5.12 which are all annotated as *interferon*-related. These three gene sets exhibit similar dynamics (see Fig 4). Such a timely upregulation was expected, as it is associated to the viral rebound after treatment interruption and was previously reported [1, 2]. The gene set M3.4 is also linked with *antiviral response*.



**Table 1. Number of significant units in DALIA-1 at a FDR of 5%.**

	Pre-ATI	Post-ATI	Units
Gene-by-gene	0	3,389	probes <sup>a</sup>
GSEA for time series	0	67	genes set <sup>b</sup>
CAMERA	0	2	genes set <sup>b</sup>
TcGSA linear	23	203	gene sets <sup>b</sup>
TcGSA cubic polynom.	69	216	gene sets <sup>b</sup>
TcGSA splines	68	219	gene sets <sup>b</sup>

<sup>a</sup> 32,978 probes investigated after filtering

<sup>b</sup> 260 immune-related gene sets investigated (29 gene sets were automatically discarded because less than 10 probes were observed)

doi:10.1371/journal.pcbi.1004310.t001

## Another application: Influenza and pneumococcal vaccines responses

In a recent paper, Obermoser et al. [6] investigated the response to influenza and pneumococcal vaccines in healthy individuals at the gene expression level.

**Study design.** Healthy, young adults were randomly split in three groups of six volunteers each, receiving either a 2009–2010 seasonal influenza vaccine (Fluzone), a 23-valent pneumococcal vaccine (Pneumovax23), or a placebo (saline injections). Blood samples were collected at days -7, 0, 1, 3, 7, 10, 14, 21, and 28 to measure gene expression in whole blood. A more detailed description of the study can be found in Obermoser et al. [6].

**Original analysis.** In their modular analysis, Obermoser et al. [6] focused on 62 of the 260 available gene sets defined in Chaussabel et al. [13]. They investigated the changes of gene expression in those 62 gene sets for each of the seven time points from day 1 to day 28 in regards of the baseline, that was considered as the average of the two measurements at days -7 and 0. So hierarchical structure of the data was not taken into account. The three arms (saline, flu and pneumococcal) were analyzed separately, and only significant gene sets at day 1 and day 7 (not further on) are presented in their paper. Changes in eight gene sets were common to both vaccines: M4.6 (inflammation), M6.6 and M6.13 (apoptosis/cell death) and modules M4.1 and M4.15 (T cells), M4.3 (protein synthesis), M5.11, and M6.9 (no functional annotation). Nine gene sets were uniquely changing after the influenza vaccine, three were associated with antiviral responses (M1.2, M3.4, M5.12) and included genes coding for interferon (IFN)-inducible molecules. Six gene sets were uniquely responsive to the pneumococcal vaccine. Of these, five were modules including genes associated with inflammation: M3.2, M4.2, M4.13, M5.1 and M5.7.

**TcGSA results.** To compare the gene expression at the gene set level between the vaccine arm (flu or pneumococcal) and the placebo (saline) arm, we applied TcGSA on these data using Eq (2) (for each vaccine separately). In both vaccines, a large response is observed at Day 1. To avoid smoothing down the expression at  $t_i = 1$ , we used the following function of time to model the dynamic evolution of gene expression:

$$f_m(t_i) = (\eta_m + h_g)\mathbb{1}_{\{t_i=1\}} + (\eta'_m + h'_g)\mathbb{1}_{\{t_i \neq 1\}}$$

with  $(h_g, h'_g) \sim \mathcal{N}(0, \Sigma_h)$ , and  $m$  the group (either vaccine or placebo).

Most of the 62 investigated gene sets presented a significantly different evolution in vaccine arms compared to the placebo arm. Globally, the intensity of the response was stronger with

the pneumococcal vaccine than with the flu vaccine (Fig 5). The early response induced by the pneumococcal vaccine was dominated by inflammation whereas the top signal triggered by the flu vaccine involved an interferon signature (Fig 5B and 5D). In both vaccine, a T-cell response was also visible. In the pneumococcal vaccine, a plasma cell signal, in association with cell cycle gene sets (Figs 5A and 5C), started at Day 7 until Day 14. This plasma blast signal was much less clear in the flu vaccine (Figs 5B and 5D). This is in agreement with the results of Obermoser et al. modular analysis.

TcGSA offers an extended and appropriate hierarchical analysis of these data. It provides a truly longitudinal insight into the vaccine responses, that are intrinsically compared to the placebo response. One of the main difference from the results presented in Obermoser et al. paper [6] is that, according to our analysis, the inflammation gene sets (M3.2, M4.13, M5.1 and M5.7) were also involved with the flu vaccine and were not specific to the pneumococcal vaccine. This result is important as it means that both vaccine involved these inflammatory pathways. This result was not obvious from the original analysis because their approach was less powerful compared to TcGSA.

## Assessment of statistical performances on simulated data

In order to assess the behavior of the proposed method, a simulation study of TcGSA has been performed. The simulation scheme was chosen to be very close to the motivating example: the DALIA-1 trial. In each simulation run, gene expression data was simulated for 20 patients at 8 time points. 5,000 genes were simulated, divided into 100 non overlapping gene sets of 50 genes each. Each of the 100 gene sets was either simulated under ( $H_0$ ) or ( $H_1$ ). The proportion of genes under ( $H_1$ ) varied between 0%, 27% (which corresponds to results found in pre-ATI) and 85% (which corresponds to results found in post-ATI). When there are gene sets simulated under ( $H_1$ ), 75% of those were homogeneous (simulated with parameters close to those estimated for gene set M1.2 in DALIA post-ATI—see Fig 4) while the remaining 25% were heterogeneous (simulated with parameters close to those estimated for gene set M7.1 in DALIA post-ATI—see Fig 4).

Statistical performances of the proposed method are presented in Table 2. Without correcting for the fact that 100 gene set were investigated by TcGSA at each simulation runs, the average Type I error (the probability of rejecting  $H_0$  given that  $H_0$  is actually true) over a hundred runs was between 0.03 and 0.07 depending on the situation. But as soon as a control of the FDR was used, the Type-I error rate dropped well below 1%, regardless of the flexibility of the time function estimated (linear or cubic polynomials). The average statistical power (the probability of rejecting  $H_0$  given that  $H_0$  is actually false) is very good, always above 0.8 (dropping a little bit after multiple testing correction as expected).

Three other methods were also evaluated on those simulations, namely globalANCOVA [29] (using either permutations or an approximation to compute p-values), GSEA for time series and CAMERA. Their statistical performances are also presented in Table 2. Type I error is always well controlled by those three methods. However, both CAMERA and GSEA for time series exhibit very low statistical power (as low as 10 times less than TcGSA after multiple testing correction when there is a high proportion of significant gene sets). globalANCOVA, whose global null hypothesis is not so different from the one tested in TcGSA, performs quite well in terms of statistical power. Nonetheless it is still about 10% below TcGSA performances.

Those simulation results confirm that the higher number of selected gene sets by TcGSA in the two real-life examples presented in this paper are mainly due to the increased power of gene set analysis over gene-by-gene analysis (when repeated structure of the measurement is properly accounted for), and not to a large number of false positives.

**Table 2. Assessment of statistical performances through a simulation study.**

Percentage of simulated gene sets under $H_1$	Method	Type I error	Type I error after MTC*	Statistical power	Statistical power after MTC*
0%	TcGSA (linear)	0.0394	0.0002	-	-
0%	TcGSA (cubic)	0.0649	0.0004	-	-
0%	globalANCOVA (perm)	0.0483	0.0001	-	-
0%	globalANCOVA (approx)	0.0006	0	-	-
0%	GSEA for time series	0	0	-	-
0%	CAMERA	0	0	-	-
27%	TcGSA (linear)	-	-	0.883	0.829
27%	TcGSA (cubic)	-	-	0.882	0.810
27%	globalANCOVA (perm)	-	-	0.787	0.706
27%	globalANCOVA (approx)	-	-	0.660	0.510
27%	GSEA for time series	-	-	0.459	0.214
27%	CAMERA	-	-	0.374	0.109
85%	TcGSA (linear)	-	-	0.885	0.847
85%	TcGSA (cubic)	-	-	0.882	0.833
85%	globalANCOVA (perm)	-	-	0.785	0.728
85%	globalANCOVA (approx)	-	-	0.660	0.549
85%	GSEA for time series	-	-	0.289	0.074
85%	CAMERA	-	-	0.177	0.013

\* Multiple Testing Correction: performed via Benjamini-Yekutieli procedure with a 5% threshold.

In each simulation, 100 gene sets are simulated and significance level  $\alpha = 5\%$  is applied. This table displays the Type I error and the statistical power means over a hundred simulation runs for 3 different situations (0%, 27% and 85% of simulated gene sets are simulated under  $H_1$ ). Whenever the percentage of gene sets simulated under  $H_1$  is not null, 25% of the gene sets simulated under  $H_1$  are heterogeneous, the remaining 75% being homogeneous. Type I error is the probability of rejecting  $H_0$  given that  $H_0$  is true, i.e. for declaring a gene set significant when it actually is not. Statistical power is the probability of rejecting  $H_0$  given that  $H_1$  is true, i.e. for declaring a gene set significant when it actually is. Four methods are evaluated: i) TcGSA, the proposed approach, fitted either with a linear or with a cubic function of time; ii) the GlobalANCOVA procedure [29] in which p-values are either computed by permutation (10,000) or approximated; iii) the GSEA for time series [14], iv) CAMERA [31]. Default values are used for the various methods (see [S1 Text](#) and [S1 Software](#)).

doi:10.1371/journal.pcbi.1004310.t002

## Discussion

In this paper, we present a method to analyze repeated measurements of gene expression using a gene set approach. Provided that the definition of the gene sets is relevant, this method helps with detecting and interpreting subtle changes of gene expression over time. In our applications where the same definition of gene sets has been applied, we were able to compare the response to several vaccines (against HIV, Influenza and Pneumococcus). Interestingly, we found common pathways that were triggered by all three vaccines, mostly related to inflammation, as well as pathways specific to each vaccine.

The capacity of the proposed approach to detect subtle changes of gene expression is due to two main factors: i) the use of a predefined gene sets that are functionally related ii) the use of all available information, taking advantage of repeated measurements using mixed models. Measurements of gene expression data in longitudinal studies may be missing because of missed visits or poor quality of the samples, leading to unbalanced data. Missing at random

(MAR) processes (i.e. when the probability of missing data is associated to the previously measured information) may lead to biased estimates when using least squares or generalized estimating equations [37]. TcGSA can cope with such issues because of the use of Maximum Likelihood to estimate the parameters of the mixed models. This is an advantage of the TcGSA approach over those of Hummel et al. [29] or Nueda et al. [32].

An increasing number of gene sets databases are available, such as KEGG [11], Gene Ontology [12], Modules [13]. An immune related subset of Gene Ontology as well as an immune related subset of KEGG pathway have been used in additional analyses (see S3 Fig). The choice of the database used for the analysis impacts the interpretability but also the limitations of TcGSA. The efficiency of TcGSA will vary according to the number of genes represented in each gene set. The size of a given gene set has an impact on its significance, as the more genes it includes, the more likely a significant variation will be detected. The average size of the Chausabel's V2 modules is 55 genes. 17% of the 260 modules include more than 100 genes, and 31% less than 20 genes. For small gene sets, the normality assumptions of random effects of the Eqs (1), (1bis) and (2) are questionable (even though inference should still be robust [59]). Nevertheless, even though we expect that the models could be misspecified in many cases (if not all), the objective of such an analysis is to detect any significant variation over time (in the spirit of Shahbaba et al. [19], a significant variability of the trajectories between the genes inside a gene set indicates a change over time regardless of the fixed effects specification). The use of flexible time functions may help to get a better fit of dynamics although beyond cubic polynomials it did not have a substantial impact in our motivating example—see Table 1. These results vary according to the dataset and the number of time points available, and we recommend to try several models to check the robustness of the results.

TcGSA focuses on self-contained hypothesis. While competitive gene sets tests are appropriate in the presence of a strong signal (e.g. when comparing pre-ATI versus post-ATI), self-contained approaches are able to detect more subtle and sparse signals (e.g. in pre-ATI). The choice between a competitive or a self-contained approach is therefore highly dependent on the biological question.

Several extensions of the TcGSA are possible for its use in other contexts. One can also model time trends with a random effect grouped on the patient level as of  $\gamma_{p,\cdot}$  instead of on the gene level as in Eqs (1), (1bis) and (2). This identifies gene sets whose dynamic differs across the patients. This option is also implemented in the TcGSA R package. TcGSA could easily be adapted to mRNA counts data. In that case, generalized linear mixed effects models could be used, with a Poisson distribution for instance, instead of linear mixed effect models that rely on a Gaussian assumption. Or else, precision weights estimated with voom [60] could also be used to model the mean-variance relationship while fitting the proposed linear mixed effect directly on log-counts data from RNA-seq, thus taking into account heteroscedasticity.

In conclusion, the method presented gives a solution for the full exploitation of any repeated measurements of gene expression data based on a gene set analysis where a great sensibility to detect subtle change, while controlling false discovery, is needed.

## Supporting Information

**S1 Fig. Density plot for both the 100,000 simulations under the null and a 100,000 sample of the corresponding  $\chi^2$  mixture approximation.**

(EPS)

**S2 Fig. Quantile-quantile plot comparing the 100,000 simulations under the null to a 100,000 sample of the corresponding  $\chi^2$  mixture approximation.**

(TIFF)

**S3 Fig. Comparison of TcGSA results on DALIA-1 trial data for the three gene sets databases considered.** During pre-ATI, 3 out of 75 gene sets were significant in the subset of KEGG, and 0 out of 131 in the subset of GO. During post-ATI, 73 out of 75 gene sets were significant in the subset of KEGG, and 101 out of 131 in the subset of GO. 2 gene sets the subset of KEGG and 20 from the subset of GO were automatically discarded because less than 10 probes or more than 500 probes were observed.

(TIFF)

**S1 Table. Selected KEGG pathways for investigating DALIA-1.** Hand picked KEGG pathways [11] of interest for investigating DALIA-1 trial data.

(PDF)

**S2 Table. Selected GO pathways for investigating DALIA-1.** Hand picked Gene Ontology pathways [12] of interest for investigating DALIA-1 trial data.

(PDF)

**S1 Text. Supporting Information about TcGSA.**

(PDF)

**S1 Software. R files helping to reproduce the results from this article.**

(ZIP)

**S1 File. The raw p-value and the adjusted false discovery rate—calculated with the Benjamini-Yekutieli procedure—in pre-ATI in the DALIA-1 trial.**

(TXT)

**S2 File. The raw p-value and the adjusted false discovery rate—calculated with the Benjamini-Yekutieli procedure—in post-ATI in the DALIA-1 trial.**

(TXT)

**S3 File. The raw p-value and the adjusted false discovery rate—calculated with the Benjamini-Yekutieli procedure—for the pneumococcal vaccine.**

(TXT)

**S4 File. The raw p-value and the adjusted false discovery rate—calculated with the Benjamini-Yekutieli procedure—for the flu vaccine.**

(TXT)

## Acknowledgments

The authors wish to thank all participants of the DALIA-1 trial and the members of the trial committees and of the ANRS/VRI study team. Trial Committee: Jacques Banchereau, Geneviève Chêne, Carson Harrod, Christine Lacabartz, Yves Levy, Monica Montes, Karolina Palucka, Laura Richert, Louis Sloan. DSMB members: William Duncan, Roy M. Gulick, Daniel R. Kuritzkes, James Neaton, Richard Pollard (Chair). Event Validation Committee members: Joseph Fay, Ronald Mitsuyasu, Jean-Paul Viard. Trial management: Derek Blankenship, Céline Boucherie, David Jutras, Bryan King, Sophie Pérusat, Elisa Priest, Charlie Quinn, Anna Laura Ross, Mathieu Surenaud. Vaccine/GMP team: Susan Burkeholder, Amanda Cobb, Charles McWilliams, Jennifer Finholt-Perry, Lee Roberts.

Computer time for this study was partly provided by the computing facilities MCIA (Mésocentre de Calcul Intensif Aquitain) of the Université de Bordeaux and of the Université de Pau et des Pays de l'Adour.

## Author Contributions

Conceived and designed the experiments: BPH JS RT. Performed the experiments: BPH JS RT. Analyzed the data: BPH JS RT. Contributed reagents/materials/analysis tools: BPH JS RT. Wrote the paper: BPH RT.

## References

1. Bécavin C, Tchitchek N, Mints-Eya C, Lesne A, Benecke A. Improving the efficiency of multidimensional scaling in the analysis of high-dimensional data using singular value decomposition. *Bioinformatics*. 2011; 27(10):1413–1421. doi: [10.1093/bioinformatics/btr143](https://doi.org/10.1093/bioinformatics/btr143) PMID: [21421551](https://pubmed.ncbi.nlm.nih.gov/21421551/)
2. Bosinger SE, Jacquelin B, Benecke A, Silvestri G, Müller-Trutwin M. Systems biology of natural simian immunodeficiency virus infections. *Current opinion in HIV and AIDS*. 2012; 7(1):71–78. doi: [10.1097/COH.0b013e32834dde01](https://doi.org/10.1097/COH.0b013e32834dde01) PMID: [22134342](https://pubmed.ncbi.nlm.nih.gov/22134342/)
3. Querec TD, Akondy RS, Lee EK, Cao W, Nakaya HI, Teuwen D, et al. Systems biology approach predicts immunogenicity of the yellow fever vaccine in humans. *Nature Immunology*. 2008; 10(1):116–125. doi: [10.1038/ni.1688](https://doi.org/10.1038/ni.1688) PMID: [19029902](https://pubmed.ncbi.nlm.nih.gov/19029902/)
4. Palermo RE, Patterson LJ, Aicher LD, Korth MJ, Robert-Guroff M, Katze MG. Genomic Analysis Reveals Pre-and Postchallenge Differences in a Rhesus Macaque AIDS Vaccine Trial: Insights into Mechanisms of Vaccine Efficacy. *Journal of Virology*. 2011; 85(2):1099–1116. doi: [10.1128/JVI.01522-10](https://doi.org/10.1128/JVI.01522-10) PMID: [21068249](https://pubmed.ncbi.nlm.nih.gov/21068249/)
5. Lévy Y, Thiébaud R, Montes M, Lacabaratz C, Sloan L, King B, et al. Dendritic cell-based therapeutic vaccine elicits polyfunctional HIV-specific T-cell immunity associated with control of viral load. *European journal of immunology*. 2014; 44(9):2802–10. doi: [10.1002/eji.201344433](https://doi.org/10.1002/eji.201344433) PMID: [25042008](https://pubmed.ncbi.nlm.nih.gov/25042008/)
6. Obermoser G, Presnell S, Domico K, Xu H, Wang Y, Anguiano E, et al. Systems Scale Interactive Exploration Reveals Quantitative and Qualitative Differences in Response to Influenza and Pneumococcal Vaccines. *Immunity*. 2013; 38(4):831–844. doi: [10.1016/j.immuni.2012.12.008](https://doi.org/10.1016/j.immuni.2012.12.008) PMID: [23601689](https://pubmed.ncbi.nlm.nih.gov/23601689/)
7. Storey JD, Xiao W, Leek JT, Tompkins RG, Davis RW. Significance analysis of time course microarray experiments. *Proceedings of the National Academy of Sciences of the United States of America*. 2005; 102(36):12837–12842. doi: [10.1073/pnas.0504609102](https://doi.org/10.1073/pnas.0504609102) PMID: [16141318](https://pubmed.ncbi.nlm.nih.gov/16141318/)
8. Berk M, Hemingway C, Levin M, Montana G. Longitudinal Analysis of Gene Expression Profiles Using Functional Mixed-Effects Models. In: Di Ciaccio A, Coli M, Angulo Ibanez JM, editors. *Advanced Statistical Methods for the Analysis of Large Data-Sets*. Springer; 2012. p. 57–67.
9. Liquet B, Le Cao KA, Hocini H, Thiébaud R. A novel approach for biomarker selection and the integration of repeated measures experiments from two assays. *BMC bioinformatics*. 2012; 13(1):325. doi: [10.1186/1471-2105-13-325](https://doi.org/10.1186/1471-2105-13-325) PMID: [23216942](https://pubmed.ncbi.nlm.nih.gov/23216942/)
10. Wang L, Chen X, Wolfinger RD, Franklin JL, Coffey RJ, Zhang B. A unified mixed effects model for gene set analysis of time course microarray experiments. *Statistical Applications in Genetics and Molecular Biology*. 2009; 8(1):47. doi: [10.2202/1544-6115.1484](https://doi.org/10.2202/1544-6115.1484)
11. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*. 2000; 28(1):27–30. doi: [10.1093/nar/28.1.27](https://doi.org/10.1093/nar/28.1.27) PMID: [10592173](https://pubmed.ncbi.nlm.nih.gov/10592173/)
12. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*. 2000; 25(1):25–29. doi: [10.1038/75556](https://doi.org/10.1038/75556) PMID: [10802651](https://pubmed.ncbi.nlm.nih.gov/10802651/)
13. Chaussabel D, Quinn C, Shen J, Patel P, Glaser C, Baldwin N, et al. A modular analysis framework for blood genomics studies: application to systemic lupus erythematosus. *Immunity*. 2008; 29(1):150–164. Available from: [http://www.biiir.net/public\\_wikis/module\\_annotation/V2\\_Trial\\_8\\_Modules](http://www.biiir.net/public_wikis/module_annotation/V2_Trial_8_Modules). doi: [10.1016/j.immuni.2008.05.012](https://doi.org/10.1016/j.immuni.2008.05.012) PMID: [18631455](https://pubmed.ncbi.nlm.nih.gov/18631455/)
14. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*. 2005; 102(43):15545–15550. doi: [10.1073/pnas.0506580102](https://doi.org/10.1073/pnas.0506580102) PMID: [16199517](https://pubmed.ncbi.nlm.nih.gov/16199517/)
15. Efron B, Tibshirani R. On testing the significance of sets of genes. *Annals of Applied Statistics*. 2007; 1(1):107–129. doi: [10.1214/07-AOAS101](https://doi.org/10.1214/07-AOAS101)
16. Maciejewski H. Gene set analysis methods: statistical models and methodological differences. *Briefings in bioinformatics*. 2014; 15(4):504–518. doi: [10.1093/bib/bbt002](https://doi.org/10.1093/bib/bbt002) PMID: [23413432](https://pubmed.ncbi.nlm.nih.gov/23413432/)
17. Liu J, Hughes-Oliver JM, Menius JA. Domain-enhanced analysis of microarray data using GO annotations. *Bioinformatics*. 2007; 23(10):1225–1234. doi: [10.1093/bioinformatics/btm092](https://doi.org/10.1093/bioinformatics/btm092) PMID: [17379692](https://pubmed.ncbi.nlm.nih.gov/17379692/)



18. Nueda MJ, Sebastián P, Tarazona S, García-García F, Dopazo J, Ferrer A, et al. Functional assessment of time course microarray data. *BMC bioinformatics*. 2009; 10 Suppl 6:S9. doi: [10.1186/1471-2105-10-S6-S9](https://doi.org/10.1186/1471-2105-10-S6-S9) PMID: [19534758](https://pubmed.ncbi.nlm.nih.gov/19534758/)
19. Shahbaba B, Tibshirani R, Shachaf CM, Plevritis SK. Bayesian gene set analysis for identifying significant biological pathways. *Journal of the Royal Statistical Society Series C, Applied statistics*. 2011; 60(4):541–557. doi: [10.1111/j.1467-9876.2011.00765.x](https://doi.org/10.1111/j.1467-9876.2011.00765.x) PMID: [21857748](https://pubmed.ncbi.nlm.nih.gov/21857748/)
20. Ackermann M, Strimmer K. A general modular framework for gene set enrichment analysis. *BMC bioinformatics*. 2009; 10:47. doi: [10.1186/1471-2105-10-47](https://doi.org/10.1186/1471-2105-10-47) PMID: [19192285](https://pubmed.ncbi.nlm.nih.gov/19192285/)
21. Prieto C, Rivas MJ, Sánchez JM, López-Fidalgo J, De Las Rivas J. Algorithm to find gene expression profiles of deregulation and identify families of disease-altered genes. *Bioinformatics (Oxford, England)*. 2006; 22(9):1103–1110. doi: [10.1093/bioinformatics/btl053](https://doi.org/10.1093/bioinformatics/btl053)
22. Hu Y, Gao L, Shi K, Chiu DKY. Detection of deregulated modules using deregulatory linked path. *PloS one*. 2013; 8(7):e70412. doi: [10.1371/journal.pone.0070412](https://doi.org/10.1371/journal.pone.0070412) PMID: [23894653](https://pubmed.ncbi.nlm.nih.gov/23894653/)
23. Conesa A, Nueda MJ, Ferrer A, Talón M. maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments. *Bioinformatics*. 2006; 22(9):1096–1102. doi: [10.1093/bioinformatics/btl056](https://doi.org/10.1093/bioinformatics/btl056) PMID: [16481333](https://pubmed.ncbi.nlm.nih.gov/16481333/)
24. Guo X, Qi H, Verfaillie CM, Pan W. Statistical significance analysis of longitudinal gene expression data. *Bioinformatics*. 2003; 19(13):1628–1635. doi: [10.1093/bioinformatics/btg206](https://doi.org/10.1093/bioinformatics/btg206) PMID: [12967958](https://pubmed.ncbi.nlm.nih.gov/12967958/)
25. Rajicic N, Cuschieri J, Finkelstein DM, Miller-Graziano CL, Hayden D, Moldawer LL, et al. Identification and interpretation of longitudinal gene expression changes in trauma. *PloS one*. 2010; 5(12):e14380. doi: [10.1371/journal.pone.0014380](https://doi.org/10.1371/journal.pone.0014380) PMID: [21187951](https://pubmed.ncbi.nlm.nih.gov/21187951/)
26. Park T, Yi SG, Lee S, Lee SY, Yoo DH, Ahn JI, et al. Statistical tests for identifying differentially expressed genes in time-course microarray experiments. *Bioinformatics*. 2003; 19(6):694–703. doi: [10.1093/bioinformatics/btg068](https://doi.org/10.1093/bioinformatics/btg068) PMID: [12691981](https://pubmed.ncbi.nlm.nih.gov/12691981/)
27. Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, Bushel P, et al. Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of computational biology*. 2001; 8(6):625–637. doi: [10.1089/106652701753307520](https://doi.org/10.1089/106652701753307520) PMID: [11747616](https://pubmed.ncbi.nlm.nih.gov/11747616/)
28. Luan Y, Li H. Model-based methods for identifying periodically expressed genes based on time course microarray gene expression data. *Bioinformatics*. 2004; 20(3):332–339. doi: [10.1093/bioinformatics/btg413](https://doi.org/10.1093/bioinformatics/btg413) PMID: [14960459](https://pubmed.ncbi.nlm.nih.gov/14960459/)
29. Hummel M, Meister R, Mansmann U. GlobalANCOVA: exploration and assessment of gene group effects. *Bioinformatics*. 2008; 24(1):78–85. doi: [10.1093/bioinformatics/btm531](https://doi.org/10.1093/bioinformatics/btm531) PMID: [18024976](https://pubmed.ncbi.nlm.nih.gov/18024976/)
30. Zhang K, Wang H, Bathke AC, Harrar SW, Piepho HP, Deng Y. Gene set analysis for longitudinal gene expression data. *BMC bioinformatics*. 2011; 12(1):273. doi: [10.1186/1471-2105-12-273](https://doi.org/10.1186/1471-2105-12-273) PMID: [21722407](https://pubmed.ncbi.nlm.nih.gov/21722407/)
31. Wu D, Smyth GK. Camera: A competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Research*. 2012; 40(17):e133. doi: [10.1093/nar/gks461](https://doi.org/10.1093/nar/gks461) PMID: [22638577](https://pubmed.ncbi.nlm.nih.gov/22638577/)
32. Nueda MJ, Carbonell J, Medina I, Dopazo J, Conesa A. Serial Expression Analysis: a web tool for the analysis of serial gene expression data. *Nucleic acids research*. 2010; 38(Web Server issue):W239–W245. doi: [10.1093/nar/gkq488](https://doi.org/10.1093/nar/gkq488) PMID: [20525784](https://pubmed.ncbi.nlm.nih.gov/20525784/)
33. Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, Park PJ. Discovering statistically significant pathways in expression profiling studies. *Proceedings of the National Academy of Sciences of the United States of America*. 2005; 102(38):13544–13549. doi: [10.1073/pnas.0506577102](https://doi.org/10.1073/pnas.0506577102) PMID: [16174746](https://pubmed.ncbi.nlm.nih.gov/16174746/)
34. Goeman JJ, Bühlmann P. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*. 2007; 23(8):980–987. doi: [10.1093/bioinformatics/btm051](https://doi.org/10.1093/bioinformatics/btm051) PMID: [17303618](https://pubmed.ncbi.nlm.nih.gov/17303618/)
35. Emmert-Streib F, Glazko GV. Pathway analysis of expression data: deciphering functional building blocks of complex diseases. *PLoS computational biology*. 2011; 7(5):e1002053. doi: [10.1371/journal.pcbi.1002053](https://doi.org/10.1371/journal.pcbi.1002053) PMID: [21637797](https://pubmed.ncbi.nlm.nih.gov/21637797/)
36. Laird N, Ware J. Random-effects models for longitudinal data. *Biometrics*. 1982; 38(4):963–974. doi: [10.2307/2529876](https://doi.org/10.2307/2529876) PMID: [7168798](https://pubmed.ncbi.nlm.nih.gov/7168798/)
37. Diggle P, Heagerty P, Liang KY, Zeger S. *Analysis of longitudinal data*. Oxford University Press, USA; 2002.
38. Berry MPR, Graham CM, McNab FW, Xu Z, Bloch Saa, Oni T, et al. An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis. *Nature*. 2010; 466(7309):973–977. doi: [10.1038/nature09247](https://doi.org/10.1038/nature09247) PMID: [20725040](https://pubmed.ncbi.nlm.nih.gov/20725040/)
39. Zak DE, Andersen-Nissen E, Peterson ER, Sato A, Hamilton MK, Borgerding J, et al. Merck Ad5/HIV induces broad innate immune activation that predicts CD8+ T-cell responses but is attenuated by

- preexisting Ad5 immunity. *Proceedings of the National Academy of Sciences of the United States of America*. 2012; 109(50):E3503–E3512. doi: [10.1073/pnas.1208972109](https://doi.org/10.1073/pnas.1208972109) PMID: [23151505](https://pubmed.ncbi.nlm.nih.gov/23151505/)
40. Doering TA, Crawford A, Angelosanto JM, Paley MA, Ziegler CG, Wherry EJ. Network analysis reveals centrally connected genes and pathways involved in CD8+ T cell exhaustion versus memory. *Immunity*. 2012; 37(6):1130–1144. doi: [10.1016/j.immuni.2012.08.021](https://doi.org/10.1016/j.immuni.2012.08.021) PMID: [23159438](https://pubmed.ncbi.nlm.nih.gov/23159438/)
41. Simonini G, Xu Z, Caputo R, De Libero C, Pagnini I, Pascual V, et al. Clinical and transcriptional response to the long-acting interleukin-1 blocker canakinumab in Blau syndrome-related uveitis. *Arthritis and rheumatism*. 2013; 65(2):513–518. doi: [10.1002/art.37776](https://doi.org/10.1002/art.37776) PMID: [23124805](https://pubmed.ncbi.nlm.nih.gov/23124805/)
42. Cliff JM, Lee JS, Constantinou N, Cho JE, Clark TG, Ronacher K, et al. Distinct phases of blood gene expression pattern through tuberculosis treatment reflect modulation of the humoral immune response. *The Journal of infectious diseases*. 2013; 207(1):18–29. doi: [10.1093/infdis/jis499](https://doi.org/10.1093/infdis/jis499) PMID: [22872737](https://pubmed.ncbi.nlm.nih.gov/22872737/)
43. Hastie T. Generalized additive models. In: Chambers, JM, Hastie, TJ, editors. *Statistical Models in S*. Chapman and Hall/CRC; 1991. p. 249–308.
44. Self SG, Liang Ky. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*. 1987; 82(398):605–610. doi: [10.1080/01621459.1987.10478472](https://doi.org/10.1080/01621459.1987.10478472)
45. Stram DO, Lee JW. Variance components testing in the longitudinal mixed effects model. *Biometrics*. 1994; 50:1171–1177. doi: [10.2307/2533455](https://doi.org/10.2307/2533455) PMID: [7786999](https://pubmed.ncbi.nlm.nih.gov/7786999/)
46. Stram DO, Lee JW. Corrections to “Variance components testing in the longitudinal mixed effects model” by D. O. Stram and J. W. Lee; 50, 1171–1177, 1994. *Biometrics*. 1995; 51(3):1196.
47. Molenberghs G, Verbeke G. Likelihood Ratio, Score, and Wald Tests in a Constrained Parameter Space. *The American Statistician*. 2007; 61(1):22–27. doi: [10.1198/000313007X171322](https://doi.org/10.1198/000313007X171322)
48. Dudoit S, Van der Laan MJ. *Multiple Testing Procedures with Applications to Genomics*. Springer Series in Statistics. Springer (New York); 2008.
49. Yekutieli D, Benjamini Y. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*. 2001; 29(4):1165–1188. doi: [10.1214/aos/1013699998](https://doi.org/10.1214/aos/1013699998)
50. Harville DA. Maximum Likelihood Approaches to Variance Component 1 Estimation and to Related Problems. *Journal of the American Statistical Association*. 1977; 72(358):320–338. doi: [10.2307/2286798](https://doi.org/10.2307/2286798)
51. Snijders TAB, Bosker RJ. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. 2nd ed. Sage; 2012.
52. Verbeke G, Molenberghs G. *Linear mixed models for longitudinal data*. Springer Series in Statistics. Springer; 2000.
53. Hitchcock DB, Booth JG, Casella G. The effect of pre-smoothing functional data on cluster analysis. *Journal of Statistical Computation and Simulation*. 2007; 77(12):1043–1055. doi: [10.1080/10629360600880684](https://doi.org/10.1080/10629360600880684)
54. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2001; 63(2):411–423. doi: [10.1111/1467-9868.00293](https://doi.org/10.1111/1467-9868.00293)
55. Xie Y, Wang X, Story M. Statistical methods of background correction for Illumina BeadArray data. *Bioinformatics*. 2009; 25(6):751–757. doi: [10.1093/bioinformatics/btp040](https://doi.org/10.1093/bioinformatics/btp040) PMID: [19193732](https://pubmed.ncbi.nlm.nih.gov/19193732/)
56. Shi W, Oshlack A, Smyth GK. Optimizing the noise versus bias trade-off for Illumina whole genome expression BeadChips. *Nucleic acids research*. 2010; 38(22):e204. doi: [10.1093/nar/gkq871](https://doi.org/10.1093/nar/gkq871) PMID: [20929874](https://pubmed.ncbi.nlm.nih.gov/20929874/)
57. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. 2007; 8(1):118–127. doi: [10.1093/biostatistics/kxj037](https://doi.org/10.1093/biostatistics/kxj037) PMID: [16632515](https://pubmed.ncbi.nlm.nih.gov/16632515/)
58. Murohashi M, Hinohara K, Kuroda M, Isagawa T, Tsuji S, Kobayashi S, et al. Gene set enrichment analysis provides insight into novel signalling pathways in breast cancer stem cells. *British journal of cancer*. 2010; 102(1):206–212. doi: [10.1038/sj.bjc.6605468](https://doi.org/10.1038/sj.bjc.6605468) PMID: [19997106](https://pubmed.ncbi.nlm.nih.gov/19997106/)
59. Jacqmin-Gadda H, Sibillot S, Proust C, Molina JM, Thiébaud R. Robustness of the linear mixed model to misspecified error distribution. *Computational Statistics and Data Analysis*. 2007; 51(10):5142–5154. doi: [10.1016/j.csda.2006.05.021](https://doi.org/10.1016/j.csda.2006.05.021)
60. Law CW, Chen Y, Shi W, Smyth GK. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome biology*. 2014; 15(2):R29. doi: [10.1186/gb-2014-15-2-r29](https://doi.org/10.1186/gb-2014-15-2-r29) PMID: [24485249](https://pubmed.ncbi.nlm.nih.gov/24485249/)